Family-Based Association Tests

and the

FBAT-toolkit

Nan M. Laird (laird@hsph.harvard.edu)

USER'S MANUAL

(Updated March 2009)

New Material Highlighted in Green

Table of Contents

1. Overview

FBAT is an acronym for Family-Based Association Tests in genetic analyses. Family-based association designs, as opposed to case-control study designs, are particularly attractive, since they test for linkage as well as association, avoid spurious associations caused by admixture of populations, and are convenient for investigators interested in refining linkage findings in family samples.

The unified approach to family-based tests of association, introduced by Rabinowitz and Laird (2000) and Laird *et al*. (2000), builds on the original TDT method (Spielman *et al.*, 1993) in which alleles transmitted to affected offspring are compared with the expected distribution of alleles among offspring.
In particular, the method puts tests of different genetic models, tests of different sampling designs, tests involving different disease phenotypes, tests with missing parents, and tests of different null hypotheses, all in the same framework. Similar in spirit to a classical TDT test, the approach compares the genotype distribution observed in the 'cases' to its expected distribution under the null hypothesis, the null hypothesis being "no linkage and no association" or "no association, in the presence of linkage". Here, the expected distribution is derived using Mendel's law of segregation and conditioning on the sufficient statistics for *any* nuisance parameters under the null. Since conditioning eliminates all nuisance parameters, the technique avoids confounding due to model misspecification as well as admixture or population stratification (Rabinowitz and Laird, 2000; Lazzeroni and Lange, 2001).

In order to adapt these "classical" family-based association tests to even more complex scenarios such as multivariate or longitudinal data sources with either binary or quantitative traits, a broader class of conditional tests has been defined (refer to Laird and Lange, 2006).

These methods have all been implemented in the FBAT-toolkit, which consists of two packages "FBAT" and "PBAT". The software provides methods for a wide range of situations that arise in family-based association studies. It provides options to test linkage or association in the presence of linkage, using marker or haplotype data, single or multiple traits. "PBAT" can compute a variety of univariate, multivariate and time-to-onset statistics for nuclear families as well as for extended pedigrees. "PBAT" can also include covariates and gene/covariate-interactions in all computed FBAT-statistics. Further, "PBAT" can be used for pre- and post-study power calculations and construction of the most powerful test statistic. For situations in which multiple traits and markers are given, "PBAT" provides screening tools to sift through a large pool of traits and markers and to select the most 'promising' combination of traits and markers thereof, while at the same time handling the multiple testing problem. For futher details on PBAT, see the PBAT webpage; the remainder of this manual will focus on the FBAT package.

Note: Throughout this document we use phenotype to denote a disease or disorder of interest. The word trait is used to refer to a specific outcome associated with the phenotype.

2. Statistical Background: Test Statistic and its Distribution.

This section may be skipped for those familiar with the theory who just want to use the package.

In this section, we briefly describe the underlying theory of FBAT statistic and its distribution, as discussed in Rabinowitz and Laird (2000) and Laird *et al*. (2000). For details on downloading the package, the input files and the coding of marker genotypes and traits, we refer to Sections 3.1-2, 5.1-3, 4.1.1.2 and 4.1.1.3

In the general approach to family-based association tests proposed by Rabinowitz and Laird (2000), tests for association are conceptualized as a two-stage procedure. The first stage involves defining a test statistic that reflects association between the trait locus and the marker locus. The second stage involves computing the distribution of the genotype marker data under the null hypothesis by treating the offspring genotype data as random, and conditioning on other aspects of the data. These two stages allow a great deal of flexibility in the construction of tests applicable in many different settings.

With complete parental data, the null distribution is obtained by conditioning on the observed traits in all family members and on the parental marker genotypes. For incomplete parental data, the null distribution is obtained, not only by conditioning on all observed traits and any observed parental marker genotypes, but also on the offspring genotype configuration. Note that any partially observed parental genotypes and the offspring genotype configuration are sufficient statistics for the missing parental genotypes (Rabinowitz and Laird, 2000). By using these conditional distributions in deriving the distribution for the test statistic under the null, biases due to population admixture or stratification, misspecification of the trait distribution, and/or selection based on trait is avoided.

2.1. Defining the Test Statistic

The general "FBAT" statistic U (Laird *et al*., 2000) is based on a linear combination of offspring genotypes and traits:

$$U = S - E[S], \qquad S = \sum_{ij} T_{ij} X_{ij}, \tag{1}$$

in which X_{ij} denotes some function of the genotype of the j-th offspring in family i at the locus being tested. It depends on the genetic model under consideration. The T_{ij} is the coded trait, depending upon possibly unknown parameters (nuisance parameters). In general, the coding for T_{ij} is specified as $Y_{ij} - \mu_{ij}$. Here, Y_{ij} denotes the observed trait of the j-th offspring in family i, and μ_{ij} is seen as an offset value. More information on

T_{ij} and the choice of an appropriate offset is given further in this section and section 4.1.1.3.

The expectation in the expression for the general FBAT statistic (1) is calculated under the null hypothesis of no association, conditioning on T_{ij} and on parental genotypes. If parental genotypes are missing, we condition on the sufficient statistics for parental genotypes. Under the same null hypothesis, U is unbiased since $E(U)=0$. Using the distribution of the offspring genotypes (treating X_{ij} as random and T_{ij} as fixed), $V = Var(U) = Var(S)$ can also be calculated under the null and used to standardize U. An explicit formula for V is given in the technical report that accompanies the "FBAT" software. If X_{ij} is a scalar summary of an individual's genotype, then the large sample test statistic

$$Z = U / \sqrt{V} \tag{2}$$

is approximately $N(0,1)$. If X_{ij} is a vector, then

$$\chi^2 = U' V^{-} U, \tag{3}$$

has an approximate χ^2-distribution with degrees of freedom equal to the rank of V. Here, V^{-} denotes the inverse of V (or a generalized inverse when the inverse does not exist; this generalized inverse is based on the singular value decomposition of V – Press *et al.* 1986)

The actual test results will differ depending upon how the user specifies T_{ij} and X_{ij}, and how the distribution of X_{ij} (hence of U) is determined. Some notes on the specification of X_{ij} and T_{ij} are given below. For application-specific definitions, we refer to Section 4. Comments on the distribution of U are given under "Test Distribution" in this section.

2.1.1. Coding the marker genotypes X

The specification of X_{ij} is determined by the genetic model under consideration and by whether one wishes to test each allele separately or to perform a multivariate test.

The gene may act on the trait in a recessive, dominant or additive way, each of which gives rise to a particular scoring system (Schaid, 1996). Alternatively, the coding may be such that each possible genotype can affect the trait in an arbitrary way (adopting so-called genotype coding). For example, with the additive model, the scalar X_{ij} counts the number of a particular allele that the *ij-th* offspring has. In the multiallelic setting, X_{ij} is a vector of the number of alleles of each type that the *ij-th* individual has. More details on coding are provided in Section 4.1.1.2.

Note that model choice does not invalidate the test under the null hypothesis, but may reduce power under the alternative. Hence, it might be instructive to perform power analyses by assuming different underlying genetic models (Section 4.3.). Several studies have shown that the additive model has good power, even when the true

genetic model is not an additive one (e.g., Knapp, 1999; Tu *et al.*, 2000; Horvath *et al.*, 2001). That is why the additive model is the default in "FBAT".

If the marker has more than two alleles or the genotype model is used, "FBAT" allows two strategies: each allele (or genotype) is tested separately, resulting in multiple, single degree-of-freedom tests, or all alleles are compared simultaneously to their null expectation in one test with multiple degrees of freedom. In this case, X_{ij} is a vector (refer to Section 4.1.1.2) and the test statistic will follow a chi-square distribution under the null.

With marker data on the sex chromosomes, the coded values for females are exactly the same as they are for autosomal chromosomes, but the values are coded differently for males, as described in 4.1.1.2.

2.1.2. Coding the traits *Y*

Recall that the notation Y_{ij} refers to the trait of the *j*-th offspring in family *i,* and that T_{ij} indicates some function of the trait Y_{ij}, depending upon possibly unknown parameters. Detailed information on recoding Y_{ij} to a trait value T_{ij} in a variety of settings is given in Section 4.1.1.3. Here, we summarize important considerations to keep in mind:

- "FBAT" can handle several types of trait values Y_{ij}, e.g., dichotomous, measured or time-to-onset. However, each type will affect the selected strategy for coding (i.e., specifying T_{ij}).

- In general, T_{ij} can be any function of the trait Y_{ij} and/or other information in the data that does not depend on offspring genotypes.

- The trait must not be a function of the marker values in order to preserve the validity of the family based association test under the null hypothesis of no association. The distribution of the test statistic U conditions on trait values T_{ij}. Traits are considered as fixed whereas the marker data are considered to be random.

- Coding strategies can be based on model assumptions (prior knowledge about the population prevalence) or can be purely statistically based (e.g., choose a coding that minimizes the variance of the test statistic under the null or maximizes its power under an assumed alternative).

- In particular, T_{ij} can be adjusted for covariates. Whereas for dichotomous traits it is probably not worthwhile to adjust for them, incorporating covariate information for measured traits may substantially reduce the variability. In this case, adjusting for covariates can make an important difference in the power of the test.

- If $T_{ij}=0$ for a subject, then this subject contributes nothing to S, $E(S)$, nor $V=Var(S)$, i. e., does not contribute to the value of the test. Such individuals only help to determine the distribution of the sibling's genotypes in the case where parents have missing genotype data. Care has to be taken in coding missing or unknown traits to ensure that any T_{ij} computed for this subject will be zero (Section 5.2.2). Any unknown parental traits should also be coded as zero for affection status in the ped file or missing (-) in the phe file.

- Sample design (e.g., trios only versus also sampling unaffecteds or using quantitative traits) influences optimal choices for T_{ij}. It is shown in Lange *et al.* (2002b) how for quantitative traits, the ascertainment scheme (e.g., total population sample versus sampling from the upper tail phenotypic distribution) may influence the effect of offset choice on the quantitative FBAT statistic. An optimal choice for T_{ij} may be a transformation of Y_{ij} ($T_{ij}=Y_{ij}- \mu_{ij}$) that maximizes the power of the test statistic (Lange *et al.*, 2002a,b).

- The power of the proposed family-based association tests can depend heavily on the selected coding (e.g., the choice of offset μ_{ij} in $T_{ij}=Y_{ij}- \mu_{ij}$), unless Y_{ij} is constant for all offspring, i.e., Y_{ij} =1 (only affected offspring are used in the test). This is especially true for quantitative traits (refer to previous item and Lange *et al.* 2002b for a discussion of different scenarios).

2.2. Test Distribution

The FBAT test statistic is based on the distribution of the offspring genotypes conditional on any trait information and on the parental genotypes. If the parental genotypes are not observed, the test statistic is conditioned on the sufficient statistics for the offspring distribution. This approach of conditioning on the trait and the parental genotypes follows the original TDT; it fits within the general framework of tests which condition on the sufficient statistics for any nuisance parameters. Because the conditional offspring genotype distribution under the null can be computed simply using Mendel's segregation laws, the FBAT approach is completely robust to model misspecification.

Currently "FBAT" handles pedigrees by breaking each pedigree into all possible nuclear families, and evaluating their contribution to the test statistic independently. "PBAT" is similar in all respects to "FBAT" except for the handling of pedigrees. "PBAT" conditions on the founder genotypes, or their sufficient statistics if they are missing, to obtain the joint distribution of all the offspring in the pedigree (Rabinowitz and Laird, 2002).

In deriving the conditional null distribution of the genotype marker data, we need to be more specific about the null hypothesis itself. Family-based tests have a composite

alternative hypothesis and consequently also a composite null hypothesis: Either the null hypothesis is "no association and no linkage" or "no association in the presence of linkage" (Laird *et al.*, 2000). It is important to distinguish between the two since they give rise to different distributions for X_{ij} under the null when there is more than one offspring in the family, or when there are several nuclear families within the pedigree.

An algorithm for computing the conditional distribution for different configurations of observed marker data, given the minimal sufficient statistic under either null hypothesis, is described in Rabinowitz and Laird (2000). It can be used to compute the expectation and variance of U under the null hypothesis. Lange *et al.* (2002a,b) extended these conditional distributions to incorporate a genetic model under the alternative hypothesis, and thus allow for power calculations.

Since under the null hypothesis of "no association and no linkage" transmissions within different nuclear families in the same pedigree are independent of each other, pedigrees can be broken into nuclear families and the separate families can be treated as independent. This is the default approach in our "FBAT"- subpackage.

Under the null hypothesis of "no association in the presence of linkage" sibling marker genotypes are correlated and nuclear (pedigree sub-) families can no longer be treated as independent. However, Lake *et al*. (2000) show that a valid association test in the presence of linkage is performed using the mean of the test statistic computed via the Rabinowitz-Laird algorithm under the null hypothesis of "no association and no linkage", by using an empirical variance-covariance estimator that adjusts for the correlation among sibling marker genotypes and for different nuclear families within a single pedigree. Our tools provide an option to calculate this empirical correction to the variance.

2.3. Remarks:

- Suppose a single trait Y_{ij} for the *ij-th* offspring given X_{ij} can be modeled by a generalized linear model with a distribution from the exponential family. Then the likelihood score is given by the U statistic (1) for an appropriate coding of the trait Y_{ij}, conditioning on the sufficient statistic for any nuisance parameter under the null hypothesis (Lunetta *et al.*, 2000). Hence, score equations are a useful device for defining test statistics in other settings.
  With this observation, the generalization of (1) to multiple traits (dichotomous or measured) and/or multiple markers is intuitively straightforward, by using a multivariate score (Lange *et al*., 2002d) based on a generalized estimating equations approach (Liang and Zeger, 1986; Heyde, 1997), where there is no need to make assumptions about the phenotypic observations.

- It should be noted that score theory applied to generalized linear models or proportional hazards models is merely a device for generating potentially

useful test statistics. Indeed, relating a mean trait to marker alleles, whether disease susceptibility alleles, marker genotypes or haplotypes, often relies on making unverifiable (model) assumptions. In addition, score theory is built on the independence criterion of responses conditional on covariates. However, in the FBAT setting this does not limit the validity of the test statistic, because the distribution of the test statistic under the null does not depend on the model assumptions underlying the score test.

Also note that the general FBAT statistics (2 - large sample *Z* statistic) or (3 - large sample χ^2 statistic) fit perfectly in the broader class of conditional tests as introduced by Lange and Laird (2002c), using a "weight" matrix whose elements depend on the parental genotypic information and on the trait vector *Y* that corresponds to *X*.

- The *Z* and χ^2 tests produced by "FBAT" are large sample tests, based on the number of informative families. In "FBAT", a family is informative when it has a non-zero contribution to the FBAT statistic. We recommend testing only when 10 or more informative families are available; if very low α-values are used, the mimimum number of informative families should be higher than 10 (refer to the *minsize* command in "FBAT") as α-values smaller than 0.05 tend to make the test conservative. Work is on the way to calculate exact *p*-values for FBAT tests. For multi-allelic tests, including haplotype tests, we recommend requiring a minimum frequency of 5% for any allele (or haplotype) to decrease the degrees-of-freedom of the thest when there are multiple alleles with small frequencies.

3. The FBAT-tools Package

This Section gives a general display of what the FBAT is capable of, with links to downloadable software.

3.1. Brief Description of "FBAT"

The "FBAT" program is both interactive and command driven and provides numerous enhancements to standard family-based association tests, such as:

- Dichotomous, measured, or time-to-onset traits may be analyzed for association. Traits may be adjusted for covariate effects. With dichotomous traits, both affected and unaffected offspring can be used. Multiple traits can be analyzed simultaneously using a multivariate option.
- The program constructs, by default, a test of the null hypothesis "no linkage and no association". The *–e option* produces a test that is valid when linkage is present and the null hypothesis is simply "no association".
- The program uses data from nuclear families, sibships, or a combination of the two, to test for association between traits and genotypes. The program can handle any number of offspring in a family.
- If data are available on pedigrees, the program decomposes each pedigree into individual nuclear families or sibships that are treated as independent, except in the calculation of an empirical variance required in deriving a valid family-based association test under the null of "no association in the presence of linkage".
- "FBAT" provides estimates of allele frequencies for each marker and checks the offspring genotypes of every nuclear family for any discrepancies from Mendelian laws.
- The program computes both biallelic tests and multiallelic tests of association. It uses standard genetic models (additive, dominant, recessive or genotype) to test association. It also allows users to implement the Sibs Disequilibrium Test (SDT - Horvath and Laird, 1998).
- Multiple markers can be read simultaneously. They are used by the *fbat* command to generate multiple family-based association tests, one per marker. Multiple tightly linked markers can be used to construct haplotypes and further analysed using the *hbat* command; *hapfreq* provides estimates of haplotype frequencies and pairwise linkage disequilibrium between the specified markers.

- Three multi-marker tests are available; none requires resolving phase and all may be used without assuming no recombination between markers.

- FBAT 2.0.2 and later versions provide capacity for analyzing sex-linked markers. Analysis using the x-chromosome data is impelemented using the same commands as for the autosomal data, except that in using the 'load' command for the ped file, the –x options should be used. In addition, in the ped file, males should be coded as homozygotes.

- FBAT now reads a mapfile which is convenient for whole genome association data. The mapfile must be loaded before the ped file.

3.2. Software Downloads and Installation Information

"FBAT"

"FBAT" is accessible via several platforms:

- MacOS9/X (carbon)
- MacOSX/Darwin
- Windows
- Sun Solaris
- Linux-x86

The "FBAT" packages for various platforms are compressed and named using the following conventions:

MacOS: fbatxxx_carbon.sit; use StuffIt Expander to expand
MacOSX/Darwin: fbatxxx_darwin.tar.gz; gzcat <filename> | tar xvf -
Window: fbatxxx_win.zip; unzip using any zip utility
Sun Solaris: fbatxxx_solaris.tar.gz; gzcat < filename> | tar xvf -
Linux/x86: fbatxxx_linux.tar.gz; gzcat <filename> | tar xvf -

where xxx is the version number.

4. Types of Analyses

This Section can be seen as an extension of Section 2, in which further details are provided on the FBAT statistics and underlying assumptions in a variety of data analysis problems. These problems include 1) testing for linkage and association with a single marker, 2) testing for linkage and association with multiple markers (including haplotypes) and 3) testing for association in the presence of linkage.

We expand on the first problem setting, since most of the comments are applicable to other testing problems as well, and conclude with some notes on power calculations.

4.1 Testing for Linkage and Association using "FBAT-tools"

In the following we give some guidelines on the particular form of the FBAT statistic that allows testing for linkage and association in the following settings:

- between a marker and a disease susceptibility locus,
- between a haplotype (or multiple tightly linked markers) and a disease susceptibility locus.

In either case, information is given about assumptions that need to be verified in order for the test results to be valid, the coding of offspring genotype information and the coding of trait information.

A discussion on the use of trait information is organized separately for the case of a single trait (either dichotomous, continuous or censored) and the case of multiple traits. Special attention is given to the choice of offset value(s) and the inclusion of covariate information.

4.1.1 *Linkage and Association Between a Single Marker and a Disease Susceptibility Locus*

4.1.1.1 *Assumptions with One or Multiple Traits Being Tested*

- The null hypothesis states there is no linkage and no association between the marker locus and any trait-influencing locus. The alternative hypothesis states there is both association and linkage. In the presence of multiple traits, the null hypothesis is phrased as no linkage and no association between the marker and any genetic locus influencing any of the selected traits. The alternative hypothesis assumes both association and linkage to at least one gene influencing one or more traits.
- The sampling frame assumes subjects to be selected on the basis of trait alone (i.e., without reference to the individual's marker alleles, Section 2.1.2 – consideration 3).
- Simple Mendelian checks are used to discard data on families who do not show Mendelian inheritance patterns.
- Because of the conditioning argument on sufficient statistics, no assumptions about the trait distribution, the genetic model and the parental genotype distribution are made.

4.1.1.2 *Details on Coding X_{ij}*

In order to use the large sample Z statistic (2), we need to choose an appropriate coding for X and Y (Section 2).

For markers on autosomal chromosomes, or for a sex-linked chromosome marker in females: A recessive coding is given for allele A_1 by setting X_{ij}=1 if the ij-th individual has genotype A_1A_1 and zero otherwise. Dominant coding is achieved in a similar fashion: it involves coding X_{ij}=1 if the offspring has *any* number of A_1 alleles, and zero otherwise. An additive coding reflects an underlying additive or multiplicative genetic model and is achieved by letting X_{ij} count the number of A_1 alleles. The additive model is the default model in "FBAT". For multiallelic tests, X_{ij} is a vector with length equal to the number of alleles. Each element of the vector codes for one of the alleles.

For a sex-linked marker in a male: A recessive coding for A_1 is given by setting X_{ij}=1 if the ij-th individual has allele A_1 and zero otherwise.

Dominant coding is achieved in a similar fashion: X_{ij}=1 if the offspring has A_1 , and zero otherwise. An additive coding is achieved by letting X_{ij} count the number of A_1 alleles (either 0 or 1).

A more thorough description of marker codings is given in Schaid *et al*. (1996).

Bi- and multiallelic markers can also be treated on the genotype level by using a so-called genotype coding. Here, the possible genotypes are treated as different "alleles" and the writing is similar to multiallelic codings, where the "number of alleles" refers to the "number of genotypes". So, for example, if A_1 and A_2 are the two possible alleles for a marker, X_{ij}=(1,0,0) if the *ij-th* person is A_1A_1, X_{ij}=(0,1,0) if the *ij-th* person is heterozygous and X_{ij}=(0,0,1) if this person is homozygous A_2A_2 . NOTE: This option (genotype coding) is not available in fbat202 or later; it is available for autosomal markers in earlier versions.

Note that for multiallelic codings, X_{ij} is a vector instead of a scalar, and therefore the statistic will be multidimensional, as in the large sample chi-square statistic (3).

4.1.1.3 *Details on Coding Y_{ij}: A Single Trait*

A single dichotomous trait

There are several ways of recoding a trait Y_{ij} into T_{ij} as in the general FBAT statistic U (1). In this section we consider traits that are a dichotomous indicator of affection status (affected or not). We restrict attention to recoded traits of the form $T_{ij} = Y_{ij} - \mu$, in which μ is seen as an offset value, and Y_{ij} is a dichotomous trait.

When the trait is dichotomous, the usual approach in family-based association testing has been to consider allele transmissions from parents to affected offspring only (Spielman *et al.*, 1993). This can be achieved by setting T_{ij} =1 for affected individuals and T_{ij} =0 for all others ("FBAT" command: *setafftrait* 1 0 0). It is the default trait coding used by "FBAT". This is equivalent to selecting a zero offset μ in $T_{ij} = Y_{ij} - \mu$. Note that the default trait uses the affection status variable defined in the pedigree input data file, recoded as 1 if affection status is 2 (i.e., affected), and zero otherwise (i.e., for unaffected or unknown affection status). The trait value for an unknown affection status should always be set to zero, regardless of how the trait is defined for the rest of the sample. Different choices of μ are discussed below.

Offset options: The choice of μ in $T_{ij} = Y_{ij} - \mu$ for a dichotomous trait

a. Disease Prevalence

The theory of score tests (Laird *et al*., 2000; Whittaker and Lewis, 1998; see also Section 2.3 - Remarks) suggests using $T_{ij} = Y_{ij} - \mu$, where μ is the disease prevalence. With a rare disease, μ is nearly zero, and the result is close to the default coding in which $T_{ij} = Y_{ij}$. For more common diseases, taking $0<\mu<1$ can increase the power of the test (Lange *et al.*, 2002a; Section 4.1). Note that choosing $0<\mu<1$ allows both affecteds and unaffecteds to contribute to the test statistic. For illustrative purposes, with μ=0.5 an affected subject has $T_{ij} = 0.5$ and an unaffected subject has T_{ij} =-0.5, i.e., they receive equal weight, but have different signs. So the statistic is a contrast of transmissions to affecteds versus unaffecteds. The problem is how to estimate μ *w*hen disease prevalence is unknown for the study population. When ascertainment depends upon Y_{ij}, a valid estimate of μ usually cannot be obtained from the sample.

b. Minimizing the Variance of the FBAT Statistic

An alternative is to choose an offset μ in such a way that the variance of U (1) is minimized (Lunetta *et al.*, 2000). With n_{Aff} (n_{Unaff}) the total number of transmissions to affected (unaffected) individuals scored, it can be shown that this *Var(U)* is minimized by μ=n_{Aff}/(n_{Aff} + n_{Unaff}). Note that this is approximately the sample prevalence, where offspring are weighted by the number of heterozygous parents. In general, this offset should not be thought of as prevalence; it is merely a device for including unaffecteds in an optimal way by minimizing the variance of the test statistic.

c. Including Covariates in the Calculation of the Test Statistic

In principle, it is straightforward to adjust the dichotomous trait for covariates, by incorporating them into the estimate of μ. Using covariates requires estimation of parameters β_0 and β_1 in the regression expression under the null (Lunetta *et al.*, 2000)

$$g(\mu_{ij}) = \beta_0 + \beta_1 Z_{ij},$$

in which $\mu_{ij} = E(Y_{ij})$ is the expectation of Y_{ij} under the null, g is an appropriate link function (e.g., logistic link) and β_1 is a vector when multiple covariates are considered. The estimates can then be used to calculate the residuals $Res_{ij} = (Y_{ij} - \mu_{ij})$, $\mu_{ij} = g^{-1}(\beta_0 + \beta_1 Z_{ij})$. Covariates are accounted for in the FBAT statistic, by using Res_{ij} instead of the original traits Y_{ij} in the expression of the test statistic. Hence, once the

user has implemented the modeling [$g(\mu_{ij}) = \beta_0 + \beta_1 Z_{ij}$] outside of "FBAT" and the residuals Res_{ij} are calculated, Res_{ij} can be entered as `new' traits in a "phenotype data file" (refer to section 5.2 on setting up a phenotype data file).

Note that the conditioning argument used in deriving the distribution of the test statistic also removes any bias due to confounders not accounted for. Including covariates in the association model for Y_{ij} is therefore not necessary, but may increase efficiency (Lunetta *et al.*, 2000). In this case, we assume that covariates are unaffected by any gene linked to the test locus described by X_{ij}. Using as a covariate a marker that is tightly linked to the test locus violates this assumption. In general, the assumption is violated when $\mu_{ij} = g^{-1} (\beta_0 + \beta_1 Z_{ij})$ is a function of X_{ij}, or when X_{ij} cannot vary independently. This may bias the test since the distribution of the test statistic conditions on observed traits. It is therefore computed assuming Y_{ij}-μ_{ij} is fixed (i.e., not a random variable), whereas X_{ij} is allowed to vary.

<u>*Software:*</u>

An single offset for T_{ij} that applies to all offspring in all families may be used in the "FBAT" software using the *setafftrait* command (for preselected value) or the *offset* command and *-o option* in *fbat*, for the offset that minimizes the variance of the test statistic.

The use of covariates in the FBAT statistic provided by the "FBAT" software, requires externally calculating the residuals Res_{ij} before and submitting these as traits in a phenotype data file.

A single measured trait

There are several ways of recoding quantitative information Y_{ij} into T_{ij} (T_{ij} as in (1)). Here, Y_{ij} represents a quantitative trait. One convenient choice is to define $T_{ij} = Y_{ij} - \mu_{ij}$ and to assume that all offset values are identical ($\mu_{ij} = \mu$). Other choices discussed below include ranking the Y_{ij}'s or converting them to normal scores.

<u>Using an offset; the choice of μ in $T_{ij} = Y_{ij} - \mu$ for a measured trait</u>

a. Phenotypic Mean

In a recoded trait of the form $T_{ij} = Y_{ij} - \mu$, a common approach for a quantitative trait Y_{ij} is to mean center it, motivated by score theory

(Section 2.3 - Remarks). Here, μ is simply a (weighted) sample mean of the Y_{ij}'s.

b. Minimizing the Variance of the FBAT Statistic

As with dichotomous traits, an alternative is to choose μ in such a way that the variance of U (1) is minimized. The value of μ that minimizes *Var(U)* is given by the sample average of Y_{ij}, weighting each offspring by the number of heterozygous parents (Lunetta *et al.*, 2000).

c. Including Covariates in the Calculation of the Test Statistic

Covariate information Z_{ij} can easily be incorporated by defining $T_{ij} = (Y_{ij} - \beta_0 - \beta_1 Z_{ij})$. Note that T_{ij} is a residual, the result of removing any covariate effect on Y_{ij} under the null. Hence, instead of estimating a single μ for the entire sample, many offset parameters μ_{ij} are used.

Software:

Approaches a. and b. to compute the offset are available in "FBAT". For instance, the sample average of a single trait can be assigned as offset value using the *offset* command. Note that only one value at a time can be assigned as offset value in this way. The user-specific definition for T_{ij} via minimizing *Var(U)* can be invoked by the "FBAT" *-o option* within the *fbat* command. Since the "FBAT" software puts μ=0 by default, phenotypic information should be submitted to the package in re-coded form, i.e., as T_{ij} instead of Y_{ij}, unless μ is to be calculated post hoc to minimize the variance of U.

Environmental correlation with multiple sibs in a family

Environmental correlation between the traits of siblings in the same family can arise as a result of many shared factors beyond the genes under study. Accounting for this environmental correlation when there are multiple siblings in the family can lead to increased power for association using FBAT. Using the variance model of Fulker *et al.* (1999), the original FBAT statistic U (1) can be generalized to incorporate complex within-family structures (Lange *et al.*, 2002b); the phenotypic variance W_i for the *i-th* family may have components that are attributable to the putative quantitative trait locus, shared environmental, and polygenic effects.

The FBAT test adjusting for environmental correlation uses expression (1) for the general FBAT statistic U, but T_{ij} is replaced by the corresponding element of T_i, a vector of traits for the i-th family defined by

$$T_i = W_i^{-1}(Y_i - \mu_i).$$

Here, $W_i = \mathrm{var}(Y_i)$, Y_i is the vector of traits for the i-th family, and μ_i is the corresponding vector of offsets.

Currently, the "FBAT" software does not include options to account for shared environmental effects among the siblings of one family; this has to be done externally by creating the trait vector as $T_i = W_i^{-1}(Y_i - \mu_i)$.

Remark:

A single censored trait

For diseases of late onset, it may be helpful to model time to onset rather than affection status. In this way, unaffected subjects contribute phenotypic information in proportion to their passage through the age of risk. At other times, time to onset may be of interest for its own sake. In either case, when time-to-onset of disease/disorder is the trait, some observations may be censored. Censoring occurs when incomplete information is available on the time-to-onset survival times of some individuals. If information about censoring times is given, e.g., it is known that the true onset or survival time is larger than some observed time, power can be gained by making use of this information in the test statistic. Several suggestions have been made for censored traits. Mokliatchouk *et al.* (2001) describe a score test based on Cox regression. Horvath *et al.* (2001) use a score test based on a proportional hazards model with an exponential age-at-onset distribution. Their choice of T_{ij}, which involves a censoring indicator apart from a onset/event time or censoring time variable, is motivated by expression (1) for the general FBAT statistic U taking on the same form as the score equation of a proportional hazards model with an exponential baseline hazard function (Horvath *et al.*, 2001). This approach can be implemented with the *–c option* in "FBAT".

However, the standard logrank and Wilcoxon test statistics (Lange *et al.*, 2004) can be used to develop family-based tests of association in the presence of censored traits as well. These statistics outperform the statistic described by Horvath *et al.* (2001) (e.g., in terms of power) when the underlying distribution is not exponential and are therefore recommended

over the exponential statistic mentioned before. Moreover, the logrank approach is equivalent to the proportional hazards approach described by Mokliatchouk *et al.* (2000) when the Breslow estimator for the cumulative risk is used.

In particular, the FBAT-logrank and FBAT-Wilcoxon statistics can be described using the same formalism as the general FBAT statistic (1), now replacing T_{ij} by an appropriate function of event/censoring times Y_{ij} and censoring indicators c_{ij} (1 for event times and 0 for censoring times):

For FBAT-logrank

$$T_{ij} = c_{ij} - \hat{\lambda}(Y_{ij}),$$

for FBAT-Wilcoxon

$$T_{ij} = n_{ij}c_{ij} - \sum_{Y_k \le Y_{ij}} d_k,$$

with Y_{ij} the onset/event time of censoring time for the *j-th* offspring in the *i-th* family, c_{ij} the corresponding censoring variable (1 for event times, 0 for censoring times), $\hat{\lambda}(Y_{ij})$ the Breslow estimator (Collett, 1994) for proportional hazards models, n_k the number of offspring at risk at one of the observed event times Y_k and d_k the number of events at time Y_k.

For rules of thumb about which statistic to use in a particular situation, refer to Lange *et al.* (2004). In general, when the event time distribution seems to be skewed to the left, the FBAT-Wilcoxon is to be preferred.

Note that both FBAT-logrank and FBAT-Wilcoxon are only available in the "PBAT" software.

4.1.2 *Linkage and Association Between a Haplotype and a Disease Susceptibility Locus*

In the presence of multiple tightly linked markers, or if haplotype-specific associations are suspected, an FBAT-type test using haplotypes may be more beneficial than using the markers one by one as in Section 4.1.1.

4.1.2.1 *Assumptions*

The following assumptions coincide with those for testing linkage and association between a marker and a disease susceptibility locus.

- The null hypothesis is no linkage and no association between any of the markers used to construct the haplotypes and any gene influencing the

trait. The alternative hypothesis assumes association and linkage are both present.
- The sampling frame assumes subjects to be selected on the basis of trait alone (hence, without reference to the individual's marker alleles).
- Simple Mendelian checks are used to discard data on families who do not show Mendelian inheritance patterns.

In addition, we make the following assumptions:

- Markers used for haplotype construction are tightly linked, so that no recombination occurs between them.
- As opposed to Section 4.1.1, the conditioning approach is extended to tightly linked markers by conditioning on the sufficient statistic for resolving phase as well (Horvath *et al.*, 2004). As before, because of the conditioning argument on sufficient statistics, no assumptions about the trait distribution, the genetic model and the parental genotype distribution are made.

4.1.2.2 *Specification of the Components of the Test Statistic*

Similar in spirit to the conditioning arguments used by Rabinowitz and Laird (2000), Laird *et al.* (2000), Horvath *et al.* (2001), Horvath *et al.* (2004) suggest extending their conditioning strategy to tightly linked markers by conditioning on the sufficient statistic for resolving phase in phase-unknown parent's genotypes as well. Phase-unknown subjects can be included in the evaluation of the test statistic using a set of weights assigned to the possible phased genotypes that are consistent with any ambiguous unphased genotype G_{ij}.

More specifically, assuming offspring genotype patterns G_{ij} are phase-known, Horvath *et al.* (2004) propose the following test statistic

$$U = \sum_{ij} T_{ij}[X(G_{ij}) - E(X(G_{ij}))]. \tag{6}$$

Here, the summation is over all families i and offspring j within the i-th family, T_{ij} is the trait, and $X(G_{ij})$ is some coding of phased haplotypes for the *ij-th* offspring. Under an additive model and if $X(G_{ij})$ is a scalar, $X(G_{ij})$ counts the number of a particular haplotype that the *ij-th* offspring has. In the "multi-haplotype" setting, $X(G_{ij})$ is a vector of the number of haplotypes of each type that the *ij-th* individual has. This is exactly the form of the multiallelic FBAT statistic with a single multiallelic marker and each haplotype forming an allele. The expectation is calculated under the null and involves conditioning in all nuisance parameters, including the sufficient statistics for phase.

To incorporate cases of unphased offspring genotype G_{ij}, the genotype coding $X(G_{ij})$ in (6) may be re-defined as

$$X(G_{ij}) = \sum_k X(G_{ijk}) w_{G_{ijk}}, \tag{7}$$

where k sums over the set of possible phased genotypes that are compatible with G_{ij}. The sum of weights $w_{G_{ijk}}$ over k equals 1.

The (extended) conditional distribution of phased offspring genotype patterns is then used to calculate the expectation and variance of U under the null hypothesis of no linkage and no association. As with the single marker case, multiallelic test statistics are available, and the *mode a* command can be used to obtain both biallelic and multiallelic tests.

Remarks

- By default, "FBAT" implements the weighted version (7) of the genotype coding $X(G_{ij})$. Hence, phase-unknown subjects are accounted for in the test statistic. The weights $w_{G_{ijk}}$ are estimated by the conditional probability of observing the phased genotype G_{ijk} conditional on the observed unphased genotype. Details about the underlying EM algorithm are given in Horvath *et al*. (2004).

- When testing association in the presence of linkage and when using a sample that consists of pedigree data or multiple sibs, a robust variance estimator for U should be used (Lake *et al*., 2001). More informaton about this empirical correction is provided in Section 4.2.2.

- The validity of the test under either null hypothesis does not depend upon the choice of weights $w_{G_{ijk}}$. This is because the distribution used to evaluate the standardized statistic has a mean zero and variance one by construction under the null.

- Note that, although multiallelic tests circumvent multiple testing problems, they may lack power. A new feature in FBAT allows users to specify the minimum sample frequency for any haplotype used in the global test (minfreq). We recommend setting minfreq = .05 (in addition to using minsize to control the number of haplotypes included in the global test.

- The –p option in hbat uses the full conditional distribution of offspring haplotypes to compute an 'exact' p-value via Monte-Carlo for each haplotype separately, for the global test, and for the minimum observed p-value among the haplotypes. This option can also be used to obtain exact p-values for

single marker tests. Simulations suggest that the –p option has higher power than the asymptotic test for the global haplotype test. NEW! The –p option provides the min-p test as an alternative to the global haplotype test; may be more powerful than the global haplotype test if there is only one major haplotype associated with the phenotype. This test evaluates the statistical significance of the smallest observed p-value among all the individual haplotype scores.

4.1.3 ***Multi-marker Tests.***

FBAT has three multi-marker options which allow one to simultaneously test H0: no linkage or association between any marker and any Disease Susceptibility Loci underlying the trait. Neither test requires computing haplotypes or the assumption of no recombination between the markers. Both tests use an empirical variance-covariance matrix to estimate the covariance between the markers. Both tests can be used with the –e option and will be valid for testing H0: no association in the presence of linkage when the –e option is used.

4.1.3.1 *The Multi-marker Test: FBAT-MM*

For K markers, the multi-marker test forms a vector of K markers. The empirical variance is estimated by

$$\hat{\sigma}_{ml} = \sum_i \left\{ \sum_j T_{ij}[X_{ij}^m - E(X_{ij}^m)] \sum_j T_{ij}[X_{ij}^l - E(X_{ij}^l)] \right\}$$

where m and l denote any two markers and X_{ij}^k denotes the kth coded marker for the ijth offspring. Then the multi-marker test is given by $S_{MM} = \hat{Z}^T \Sigma^{-1} \hat{Z}$ where $\hat{Z}$ is the K vector of individual marker test statistics, and Σ is the correlation matrix computed from the σ_{ml}. See Rakovski et al., (2007) or Xu et al., (2006). Note: The multi-marker test is restricted to the additive model only.

4.1.3.2 *The Min-p test: FBAT Min-p*

This test uses Monte Carlo to obtain a p-value for the maximally significant statistic, out of the set of individual statistics. See Rakovski et al, (2007).

4.1.3.3 *The Linear Combination Test: FBAT –LC*

This test uses the screening approach described in Lange et al (2003) to form a linear combination of the individual Z statistics computed for each marker. It takes the form $Z_{LC} = \frac{\hat{Z}_\beta{}^T \hat{Z}}{\sqrt{\hat{Z}_\beta{}^T \Sigma \hat{Z}_\beta}}$ where Z_β is estimated using the conditional mean model as in Lange et al (2003). Note that FBAT-LC uses data on parents phenotypes unless they have been coded as affectation status = 0 in the ped file or as missing (-) in the phe file.

FBAT-LC may not be appropriate in heavily ascertained samples. FBAT-LC will generally be more powerful than FBAT-MM in settings where trait data are available for subjects who are not ascertained on the trait. See Xu et al.(2006). FBAT-LC can give anti-conservative results when testing multiple markers in a known linkage region.

4.1.4 *Multiple Traits*

When there are multiple correlated traits available on each offspring, it may be desirable to test them simultaneously, or using a linear combination of traits. Such test strategies can substantially reduce the power loss associated with multiple testing approaches (Lange *et al.*, 2002d).

4.1.4.1 *FBAT-GEE*

With one trait outcome per offspring, the large sample Z statistic (2) can directly be utilized to construct a χ^2-test statistic with 1 degree of freedom. Its generalization to multivariate data (with no missing phenotypic data) is straightforward by defining

$$\chi^2_{FBAT-GEE} = \widetilde{U}' V_{\widetilde{U}}^{-} \widetilde{U}, \tag{4}$$

with

$$\widetilde{U} = \sum_{ij} T_{ij} \left[X_{ij} - E(X_{ij}) \right],$$
$$V_{\widetilde{U}} = Var(\widetilde{U}) = \sum_{ij} T_{ij} T_{ij}' Var(X_{ij}).$$

Here, T_{ij} is a vector of traits ($T_{ij1}, \ldots, T_{ijk}, \ldots, T_{ijK}$) for the *j-th* offspring in family *i.* Currently "FBAT" only uses subjects have complete trait data. The mean and variance of the marker are, as always, computed under the null-hypothesis. The distribution of the genotypes is computed in exactly the same way as for a single trait. The degrees of freedom of the asymptotic chi-square distribution for $\chi^2_{FBAT\text{-}GEE}$ are given by the rank of $V_{\widetilde{U}}$, which will generally be the number of traits.

The effective use of all available phenotypic information in a single analysis is often reflected in increased power to detect the alternative. Outliers can influence the test statistic and reduce power. Hence, in the presence of outliers or when traits show highly skewed distributions, transforming the raw traits to normal scores or using ranks is recommended. More information on power issues and the definition of the test statistic can be retrieved from Lange *et al.* (2002d), in the subsequent paragraphs and in Section 4.3.

To use multiple traits in "FBAT", specify the traits of interest using the *trait* command, then use the ordinary fbat command.

Using an offset with multiple traits

The remarks of Section 4.1.1.3 concerning the offset choice for a single trait carry through in the multivariate setting. Hence a common approach for a quantitative trait *vector* $Y_{ij} = (Y_{ij1}, \ldots, Y_{ijK})$, K, the total number of traits for an offspring j in family i, is to mean-center it using the trait sample means:

$$\mu = (\mu_1, \ldots, \mu_K)', \ \mu_k = \sum_{ij} Y_{ijk} = \overline{Y}_k, \qquad k = 1, \ldots, K.$$

In this case, imputing missing traits Y_{ijk} by the observed phenotypic mean of the *k-th* trait is equivalent to setting $T_{ijk}=0$, when T_{ijk} is $Y_{ijk} - \mu_k$ and μ_k the phenotypic sample mean of the *k-th* trait. This implies that the *ij-th* individual does not contribute to the test statistic for the *k-th* trait.

Software:

As opposed to the single trait setting, "FBAT" does not allow the option to choose μ so as to minimize the variance of the statistic when multiple traits are specified. In other words, the *–o option* is not available when multiple traits are specified. However, the *–o option* can be used for each trait separately, and the adjusted trait can be calculated outside of "FBAT".
We emphasize that since the "FBAT" software sets $\mu=0$ by default, it is important to realize that phenotypic data should in general be submitted in re-coded form T_{ij} instead of Y_{ij}. So centering traits using external norms or adjusting for covariates requires the user to do so externally and to read the defined traits into the phenotype file

4.1.4.2 *FBAT-LC*

FBAT-LC for multiple traits is very similar to FBAT-LC for multiple snps, except that it combines individual test statistics across multiple traits for a single snp. As for FBAT-LC for snps, the weights are chosen by fitting the conditional mean model for each trait separately. Unlike the situation for multiple snps, an empirical variance is not necessary, and the test remains valid in the case of testing association with known linkage and multiple offspring. See (Xiao, et. al., 2007).

Software: To implement FBAT-LC for multiple SNPs, one should use the trait command to select the traits, then use fbat with the –t option.

4.2. Testing for Association in the Presence of Linkage using "FBAT-tools"

Most of the discussion of the section "Testing for Linkage and Association using FBAT" carries through. The important difference now is that there is correlation between transmissions to siblings in a family.

4.2.1 Assumptions

- The null hypothesis permits linkage between the marker and any disease susceptibility locus, but no association between haplotypes and the alleles of any trait locus linked to the haplotype locus. The alternative hypothesis assumes both association and linkage.
- Pedigrees are decomposed into individual nuclear families and are treated as independent in most of the calculations. The pedigree's contribution to the test statistic U in (1) is obtained by summing over all nuclear families within the pedigree. However, in the case where linkage is present and the null hypothesis states "no association, but linkage", the genotypes of the different nuclear families derived from one pedigree are correlated. Hence, the variance of U in the large sample *FBAT* statistic should be computed empirically without making assumptions about the recombination parameter, or the degree of correlation.

Important remark:

- When multiple sibs are present in a family, or multiple nuclear families are present in a family, and the null hypothesis assumes no association but linkage, the sibling's genotypes are correlated, with correlation depending on the linkage parameter between the marker and the unknown trait locus. The patterns of allele sharing

(identity-by-descent relationships) are not necessarily independent of the patterns of traits in the nuclear family or pedigree, and the minimal sufficient statistics used in the conditional distribution of *S* under the null should account for these identity-by-descent relationships. Thus, with multiple sibs in a family or extended pedigrees, and testing for association in an area of known linkage, the empirical estimator for *S* should be used.

- No assumptions about the trait distribution, the genetic model and the parental genotype distribution are required, because of the conditioning argument on sufficient statistics.
- Using haplotypes in association testing, the conditioning approach is extended to tightly linked markers by conditioning on the sufficient statistic for resolving the phase as well (Horvath *et al.*, 2004).
- As before, the sampling frame assumes subjects to be selected on the basis of trait alone (hence, without reference to the individual's marker alleles).
- Simple Mendelian checks are used to discard data on families who do not show Mendelian inheritance patterns.

4.2.2 Specification of the Test Statistic

The major observation is that the same general form for the test statistic as in Section 4.1 can be used when testing for association in the presence of linkage. The only difference is that the empirical variance proposed by Lake *et al.* (2000) should be used in all formulas to adjust for the correlation in transmissions to offspring.

Hence, either the univariate or the multivariate FBAT statistic can be used, with the appropriate corrections for the variance of *S*. In particular, for an empirical variance in the univariate setting, first rewrite the test statistic *S* as

$$U = \sum_k U_k,$$

where $U_k = \sum_{ij} T_{ijk}(X_{ijk} - E[X_{ijk}])$, and we let *i* index the nuclear families within a pedigree and let *k* index the pedigrees. So X_{ijk} and T_{ijk} denote the genotypes and traits for the *j*-th offspring in the *i*-th nuclear family of the *k*-th pedigree. Then the empirical variance of *U* is given by

$$Var(U) = \sum_k U_k^{\,2}.$$

If there is only one nuclear family per pedigree, the computation of the empirical variance simplifies to

$$Var(U) = \sum_i U_i^2,$$

where U_i is the sum of the i-th family's contribution to U.

This expression for var(U) is used in formula (2) for the univariate family-based association test; a straightforward extension gives the matrix Var(U) for the multivariate case. To use the empirical variance, specifiy the *option –e* for either *fbat* or *hbat* in "FBAT".

4.3 FBAT-tools and Power Calculations

Lange *et al.* (2002a,b) propose a unified approach to power calculations for family-based association tests of a single dichotomous or continuous trait and a biallelic marker. The analytical approach computes the expected conditional power of the test statistic where expectation is taken over the genotype distribution of parents and over the offspring traits, conditional on the ascertainment condition. The method is particularly attractive, since it encompasses studies using both affected and unaffected offspring, and encompasses situations with missing parental information as well. In these situations and for other designs (e.g., using multiallelic markers or multiple traits) it offers an alternative way to determine the power of family-based association tests, other than investigating power by approximation or simulation studies.

The general analytic approach to power calculations is implemented in the "PBAT" software package. Power calculations are not available in FBAT.

5. Required Input Data Files

This Section explains how to put your data into the right format for usage with the FBAT package

The data on pedigree structure, affection status, and genotype are read in from a standard pedigree file. For sex-linked data, males must be homozygous for all markers (that is, in constructing an sex-linked ped file, males should be coded as homozygous with the 'paternal' allele set to the observed maternal allele.) In addition, when loading the ped file, the –x option should be used. See instructions for the *load* command. An optional phenotype file is used for traits other than affection status. Variables are separated by a blank space or by a tab.

Note that while a continuous run of blank spaces is regarded as a single separator, each tab is treated as a separator. So there will be *k+1* fields for an entry with *k*

tabs. Generally, blank spaces as field separators are recommended. There are no line length limits.

If there are multiple nuclear families or distinct sibships in a pedigree, "FBAT" automatically breaks down each pedigree into separate sibships or nuclear families.

For whole genome data, a marker map file should be used to indicate chromosal location for the markers.

See Section 6.2 on how to load the required input files. Note that the phenotype data file is accepted by "FBAT" only after the pedigree structure has been loaded.

5.1. Pedigree Data File

The layout of the pedigree and phenotype data file shows similarities with the layout for input data files used by the LINKAGE software. Note that the latter uses additional pedigree identifiers (such as next-paternal-sibling and next-maternal-sibling numbers). Unlike the LINKAGE format, one has to specify the marker names in the first line, *unless you are using a map file. (see 5.3 below).* The pedigree data file also follows the standard pedigree file format as used by Genehunter or Mapmaker.

Layout:

First line:

names of all markers in the sequence of the genotype data

Remaining lines:

pid id fid mid sex aff A11 A12 A21 A22 …

with

Pid	pedigree ID
Id	Individual ID
Fid	father ID Use 0 (zero) for founders or marry-ins (parents not specified) in a pedigree
Mid	mother ID Use 0 (zero) for founders or marry-ins (parents not specified) in a pedigree
Sex	1 = male, 2 = female
Aff	affection status 2 = affected, 1 = unaffected, 0 = unknown
A_{ij}	allele j of marker i (j=1,2; i=1, 2,…) Alleles are represented by positive integers. Use 0 (zero) for missing alleles.

All ID's and marker names are composed of strings of any characters that do not include blank space, tab, newline, and carriage return. The maximum length for IDs and marker names are 16 and 64 characters, respectively. A maximum number of 40 alleles are allowed for each marker.

5.2. Phenotype Data File

Layout:

First line:

names of all traits in the phenotype file

Remaining lines:

pid id trait_1 trait_2 …

with *pid* and *id* the pedigree and individual ID, respectively. *Trait_i* refers to the *i-th* trait value of the individual.

Remarks:

- "FBAT" can incorporate covariate information through the definition of the offset μ_{ij} to recode the original trait Y_{ij} to $T_{ij}=Y_{ij}-\mu_{ij}$. The residuals $Y_{ij}-\mu_{ij}$, where μ_{ij} carries the covariate information, need to be calculated outside of "FBAT" (refer to section 4.1.1.3.).

Use a single hyphen (-) for missing traits. Any missing traits (-) will be recoded as zero for analysis. The order of the subject entries is not important. The set of individuals defined in the phenotype file need not be the same as that in the pedigree file (e.g., you may omit all parents in the phenotype file). However, for each individual appearing in both files, his (her) ID must be consistent. Data on any individuals in the phenotype file who do not appear in the pedigree file will be ignored.

Remark:

- Traits may be submitted in raw form Y_{ij} or in re-coded or transformed form T_{ij} (e.g., mean centered, standardized, ranked, residuals, etc.). Choice of coding has an impact on who contributes trait information to the test statistic (e.g., whether individuals with unaffected traits contribute, or to the weight given to outliers in skewed data), and has an effect on the power and efficiency of the test. Refer to the discussions about the offset choice in Section 4.1.1.).

5.3 Map File

The map file allows FBAT to distinguish markers on the sex-linked chromosomes and should be used whenever markers from the x-chromosome are in the same ped file as markers on an autosomal chromosme. The map file also facilitates the analysis of GWAS data in that it enables users to avoid reading in marker names in the first line of the ped file. When using a map file, the map file should be read in first, before the ped file, and the ped file should not have marker names in the first line.

The map file uses the following format:

marker_name chr# genetic_pos physical_pos sex_link

The count of the markers and the order in which they are listed in map the file must be the same as in the ped file.

Chr# is set to 23 for an X-chromosome marker, and 24 for Y.

Sex_link =1 if the marker is sex linked, 0 otherwise. If sex_link = 1, fbat will implement an analysis for an x-chromsome marker. For autosomal markers, and for markers in the pseudoautosomal region of the x-marker, set sex-link to zero.

As with the ped and phe files, one may use simple load command when the map File has a .map extension in the name, or one may use the load map 'filename' command.

6. A Road Map to Software Commands

Type	"FBAT" Command	Brief Description
General	? [command]	Help on specified command; "?" lists all available "FBAT" commands
	Quit	Exit "FBAT"
Data input/output	load [ped,phe,map] [-x] filename	Loading pedigree and trait data; use [-x] to load pedigree files for x-chromosome only. Use a map file for ped files with a mixture of autosomal and sex-linked markers or ped files with no marker names in first line. Load map files first.
	log [log_file,on,off]	Logging inputs and outputs
	run script_file	Running batch commands from file
Descriptive / Diagnostic	afreq [marker(s)]	Estimating allele frequencies
	genotype pedigree_id marker1 [marker2 …]	Displaying raw genotype data
	hapfreq [-d] [marker(s)]	Estimating haplotype frequencies
	viewhap [-c] [-s] [-e] [marker(s)]	Viewing haplotype configurations
	viewmarker marker [pedigree_id]	Displaying info about marker genotype distribution
Tests	displayp [p_value]	Supress p_values > specified value
	fbat [-e] [-o] [-m] [-l] [-c] [-t] [marker(s)]	Computing family-based association test statistic (FBAT)
	hbat [-c] [-e] [-o] [-p[value]] [marker(s)]	Computing haplotype version of FBAT statistic (for single markers, same as fbat command)
	Maxcmh [count]	Setting the maximum allowable number of compatible mating haplotypes
	minfreq [value]	Setting minimum frequency for alleles/haplotypes in a global test
	minsize [size_value]	Specifying minimum number of informative families
	mode [b,m,a]	Selecting biallelic, multiallelic testing procedures or both (a)
	model [a,d,r,g]	Selecting the genetic model
	offset [offset_value]	Setting the trait offset

	sdt [marker]	Computing sibship disequilibruim test (SDT)
	setafftrait aff_t unaff_t unknown_t	Setting trait values (currently unavailable)
	trait [trait_name(s)]	Specifying the trait(s) of interest

6.1. Getting Started

The general syntax for every command used in the "FBAT" program is **command [option1,option2,…]… arguments…,** where [option1,option2,…]… are optional enhancements to the basic command. Note that any command line starting with "#" will be ignored.

To display an on-line description for a specified command, use **? [command]**. If no command is specified, a listing and descriptions of all available commands is given.

Using **log [log_file,on,off]** starts logging all inputs and outputs into *log_file* or toggle the logging status. A *log_file* must be specified before you can toggle the logging status. Hence, from the moment you wish to save all inputs and outputs to a file named "session1", you first enter "log session1". The logging status will display "logging to file session1 is on". Then you may proceed with FBAT data manipulations, which will be saved to "session1". Interrupting the logging is achieved by toggling the logging status using "log off". The logging to "session1" can be resumed by entering "log on". Alternatively, a new output file "session2" is specified by the line "log session2".

All commands can be entered and stored in a text file before starting an "FBAT" session, e.g. in "script.txt". It then suffices to start the "FBAT" session and to type "**run** script.txt".

The program can be exited at any time by using **quit**.

Note that commands and options are case sensitive. A partial command name may be used to specify the command as long as it is unambiguous.

In the descriptions given below, all acceptable options for an argument are listed within a bracket and separated by commas.

6.2. Loading Input Data Files

load [ped,phe] [-x] filename

Reads in data from a pedigree or a phenotype file. The file name can be either an absolute path name or a relative path name from your current directory. The options *[ped,phe]* are not necessary when the specified *file_name* ends with a corresponding extension (*.ped* for pedigree file, *.phe* for phenotype file). Note: the –x option must be used when reading a pedfile with only x-chromososme markers. The map file (Section 5.3) must be read in first if your data file contains markers from both the X- and the autosomal chromosomes, or if your ped file does not have the marker names in the first line. Otherwise, the mapfile may be omitted.

Careful attention should be given on whether or not to recode the original quantitative traits to mean-centered traits. For additional information on recoding traits, refer to e.g., Section 4.1.1.3 (the choice of μ in $T_{ij} = Y_{ij} - \mu$). Also see Section 5.2.2 (input pedigree file).

6.3. Commands describing the marker data and its conditional distribution and commands useful for diagnostics.

afreq [marker(s)]

Outputs a sample estimate of allele frequencies for the specified marker. If no marker is specified, allele frequencies for all the markers are produced. The allele frequencies are estimated using the genotype data from parents in nuclear families only. An EM algorithm ensures that families with incomplete founder genotype data can still be incorporated in the estimation process.

Note that the estimated allele frequencies for individual markers are automatically generated via *fbat* and the *viewstat* command (see below). These estimates are not used by "FBAT" for testing purposes.

genotype pedigree-id marker1 [marker2…]

Displays the raw genotype data for the specified pedigree and marker(s). It can also be used for debugging while verifying whether the data are loaded correctly.

If father-id and mother-id refer to founders or marry-ins, their IDs will be set to zero in the display.

hapfreq [-d] [marker(s)]

Estimates haplotype frequencies of parents in nuclear families, where haplotypes are derived from the specified markers. An EM algorithm is used to resolve phase, or to include founders with missing haplotype information. The command *hbat* estimates phased genotypes, not assuming Hardy-Weinberg, and generates haplotype frequencies from the phased genotype frequencies.

Note that if only one marker is specified the allele frequencies are given similar to the *afreq* command.

The *–d option* gives pairwise values of the measures of linkage disequilibrium D for all pairs of markers, and the standardized measure of disequilibrium D' (Lewontin, 1964). Both D and D' are based on the allele frequencies. In particular,

$$D_{AB} = p_{AB} - p_{A\cdot}.p_{\cdot B},$$

where p_{AB} is the joint probability of two alleles, A and B, from two different markers and $p_{A\cdot}$ and $p_{\cdot B}$ are the marginal probabilities of alleles A and B. Also,

$$D'_{AB} = | D_{AB} | / \max(| D_{AB} |).$$

Here, $\max(|D_{AB}|)$ is computed assuming that the marginal probabilities are fixed, but p_{AB} varies.

Remark:

- In the setting of multiple tightly linked markers, the haplotype frequencies are used in computing the test statistic in order to incorporate families where there are ambiguous genotypes that cannot be resolved. Note that with a single marker, the allele frequencies are computed solely for descriptive purposes. However, the distribution of the ambiguous genotype families is computed via the conditioning algorithm, and the test statistic remains unbiased when any arbitrary set of frequencies is used. See Horvath *et al.* (2004) for details.

Note: This option is not available in the current version.

viewhap [-c] [-e] [-i] [-s] [marker(s)]

Views the haplotype configuration of the specified markers, within each family. If no marker is specified, haplotypes are constructed based on all available marker information retrieved from the pedigree file. Detailed

information about the FBAT statistics is given for the haplotypes constructed via the specified markers, including *S*, *E(S)* and *Var(S)*.

Note that if a single marker is specified, *viewhap –s marker* gives the estimated allele frequencies and the EM estimates of the phased genotype frequencies for the selected marker.

The *–i* option only prints out detailed information for informative families. If the *–s option* is specified, the information within each family is suppressed, and *S, E(S)* and *Var(S)* are given for the selected markers. The *–e option* generates empirically corrected FBAT statistics. For censored traits, the *–c option* should be used.

viewmarker marker [pedigree_id]

Displays detailed information about the marker genotype distribution (under the null of no linkage) among offspring in each nuclear family of the named pedigree (Rabinowitz and Laird, 2000).
If *pedigree_id* is not specified, the marker distributions are displayed for all nuclear families in all pedigrees.
If the family is not informative, the marker genotype distribution has probability 1 for the observed data and the output for that family is suppressed.

viewstat [-e] [-o] [-c] [-s] marker

Not available in the current version—use viewhap instead.

6.4. Testing for Linkage or Association in the Presence of Linkage

As before, all acceptable options for an argument are listed within a bracket and separated by commas.

displayp [p_value]

Selectively display the test results with *p*-value equal to or less than the specified *p*_value. If no *p*_value argument is specified, the current *p*_value is given.

The default *p*_value is 1.0 (display all results).

fbat [-e] [-o] [-m] [-l] [-c] [-t] [marker(s)]

Computes the family-based association test statistic(s) and *p*-value(s) for the specified markers (all markers -one by one- if no marker is specified) using the current trait(s) test mode (biallelic and/or multiallelic), and association model. When a single trait is considered and marker data are summarized as a scalar or a vector, the large sample *Z* statistic (2) or large sample χ^2 statistic (3) is evaluated, respectively. When multiple traits are considered, the *fbat* command calculates statistic $\chi^2_{FBAT\text{-}GEE}$ (4).

Note that pedigrees are decomposed into individual nuclear families that are treated as distinct in the calculations of the test statistic. Subjects with unknown trait (indicated in the phenotype file as hyphen '-') give a zero contribution to the test statistic.

Option -e: computes the test statistic using the empirical variance, as described in Lake *et al.* (2001). This option should be used when testing for association in an area of known linkage (the null hypothesis assumes no association but linkage) with multiple sibs in a family or when multiple families in a pedigree are used.

Option -o: uses an offset μ for the trait in the test construction (Section 4.1). With this option, the value of μ is chosen to minimize the variance of the test statistic (Lunetta *et al.*, 2000), With multiallelic genotypes, the offset is chosen to minimize the trace of the matrix *var(U)*.

This option works for both quantitative and qualitative traits. However, when this option is used for qualitative traits, trait data for both affected and unaffected offspring are required. Note that when only affected (unaffected; Y_{ij} =0) individuals are scored, $\mu=n_{Aff}/(n_{Aff}+n_{Unaff})=1$ (0) and only unaffected (affected; $Y_{ij}=1$) individuals will contribute to the test statistic (Section 4.1.1.3 – dichotomous trait). Therefore, if the data set contains only affected offspring, this option should not be used. Similarly, this option should not be used for measured traits with only a narrow range of values.

The use of *-o* does not always result in a more significant *p*-value, especially in the case where the test result is already significant using the default affection status. The *-o option* only minimizes the variance of *S* under the null hypothesis and should perform best for small departures from the null. In general, using the *-o option* will give a result similar to using the sample mean or sample prevalence to estimate μ.

Remarks:

- The *–o option* cannot be computed with *–e*, because in this situation there is no analytic expression for the empirical variance of *U*. An

approximate analysis can be done by using *–o* without *–e*, then redefining the trait outside "FBAT" as $T_{ij}=Y_{ij}-\mu_{ij}$, based on the obtained optimized offset values, and next rerunning *fbat* using the *–e option*. If Y_{ij} is dichotomous, the *setafftrait* command can be used instead.

- The *–o option* is not available when multiple traits are specified.
- The output is blank if there are no informative families.

Option –m: computes the multimarker test as described in section 4.1.3.1; see also Rakovski et al. (2006) and Xu et al. (2006). Can used in combination with –e option.

Option –l : computes the linear combination test for multiple snps as described in section 4.1.3.3 See also Xu et al. (2006). CAUTION, parental phenotypes are used in the calculation of the weights for the linear combination unless they are coded as affection status unknown (0), or as missing in a phe file. It is not advisable to use this test with subjects ascertained on outcome.

Option –p: allows computation of the min-p test statistic as described in 4.1.3.2

Option -c: allows accounting for censored observations using the maximum likelihood test statistic discussed in Horvath *et al.* (2001). We currently advocate using test statistics based on the standard logrank and Wilcoxon test instead (Lange *et al.*, 2004). The latter statistics are currently only implemented in "PBAT" .

Option –t : computes the linear combination test for multiple traits as described in section 4.1.4.2. See also Xiao et al. (2007). CAUTION, parental phenotypes are used in the calculation of the weights for the linear combination unless they are coded as affection status unknown (0), or as missing in a phe file. It is not advisable to use this test with subjects ascertained on outcome..

hbat [-c] [-e] [-o] [-p[#]] [marker(s)]

Is the haplotype version of the *fbat* command. It carries out a family-based test for association and/or linkage between the haplotype locus and any trait influencing gene.

Haplotypes are generated by the "FBAT" software on the basis of the selected markers, which are provided via a standard pedigree input file (Section 5.2.1). If no subset is specified, all markers are used to form haplotypes. When marker data is submitted to the program, all marker genotypes are treated as if unphased.

Note that when only one marker is specified, *hbat* and *fbat* give identical test results. This is not surprising, since in this case, concerns about unresolved phases are not an issue.

Options –e, -o, -c: The options *–e*, *-o* and *–c* are used in a similar fashion as within the *fbat* command. The same comments and remarks apply (see *fbat* command).

Note that if the *–o* option is used in multiallelic mode, *Var(U)* is a matrix and *–o* specifies the offset value that mimimizes the trace of *Var(U)*.

Option –p#: This option computes p-values of *Z* in the univariate case using Monte Carlo samples from the null distribution of no linkage and no association. The # specifies the number of Monte Carlo samples to be drawn; the default is 100,000. The actual number of samples may be smaller; the sampling procedure stops earlier when at least 100 Monte Carlo based *Z* values >= *Z*_obs and at least 100 Monte Carlo based *Z*_values <= *Z*_obs.

Remark:

- The *hbat* command for multivariate traits is not yet implemented.
- The *–p option* cannot be used with *–e*, since application of the latter usually implies there is linkage present. Hence, taking Monte Carlo samples, assuming independence of sibling's genotypes within a sibship, is not correct.
- The stopping rule in the *–p option* above performs well for extreme p-values. Obviously, for very small p-values under the null, the number of cycles will be large and hence a high level of precision will be obtained. Nevertheless, in the context of genome-wide screening and compounded multiple testing issues, an even higher precision level may be required. This can be achieved by increasing the default number of Monte Carlo samples to be drawn. For relatively large p-values under the null, the number of cycles will be small (i.e., the criteria of the stopping rule will be met relatively soon) resulting in a lower precision level of the p-value estimate.

maxcmh [count]

Sets the maximum allowable number of compatible mating haplotypes. The default is 1000.

minfreq [frequency_value]

Specifies the minimum allele/haplotype frequency needed to include an allele/haplotype in a multiallelic, or global haplotype test.

minsize [size_value]

Specifies the minimum number of informative families necessary to compute the test statistics. If *size_value* is not specified, the current value is displayed.
The default value is 10.

mode [b,m,a]

Specifies biallelic (*b*), multiallelic (*m*) tests or both (*a*). If no option is specified, the current mode is displayed. Currently, the *–a option* is only available with *hbat*.

The default is biallelic. For testing single SNPs, *m* and *b* modes give equivalent tests.

model [a,d,r,g]

Specifies the association model to be additive (*a*), dominant (*d*), recessive (*r*) or genotype (*g*) and can be used for both biallelic and multiallelic markers. If no option is specified, the current model is displayed. The choice of a particular model is reflected in the marker scoring scheme (Section 4.1.1.2).

The default is the additive model, since it has been shown that this model (whether dealing with biallelic or multiallelic markers) performs well (e.g., Horvath *et al*.,2001).

offset [offset_value]

Sets the offset μ to "offset_value". For a dichotomous trait, this may be the population prevalence, whenever available. For a quantitiative trait, it may refer to a sample average of the single trait. Only one value at a time

can be assigned as offset value. The command is ignored when used in the presence of multiple traits.

This command can be used with both *fbat* or *hbat*.

sdt [marker]

Computes the sibship disequilibrium test (SDT) test statistic(s) and p-value(s) for the specified marker (or all markers if no marker is specified) using the current trait and test mode (Horvath and Laird, 1998).
The SDT is designed to detect both linkage in the presence of association and association in the presence of linkage (linkage disequilibrium) when dealing with a dichotomous trait. The test does not require parental data, but requires discordant sibships with at least one affected and one unaffected sibling.

setafftrait aff_t unaff_t unknown_t

Set trait values for affected (*aff_t*), unaffected (*unaff_t*), and subjects with unknown affection status (*unknown_t*). Here affected, unaffected, and trait unknown offspring are defined using the affection status variable from the pedigree file.

Note that the trait value for unknown affection status should always remain zero. Changing the values for affected and unaffected allows unaffected subjects to contribute to the analysis. Make sure that the first two values always sum to one.

The default values are (1,0,0), i.e., only affected subjects contribute to the value of the test statistic.

trait [trait_name(s)]

Specifies the trait(s) to use for computing the test statistics. If no trait is specified, all available traits are displayed with the current trait denoted by **.
Specifying more than one trait will provide a multivariate test with multiple degrees of freedom (Lange *et al.*, 2002d).

The default trait is dichotomous (1 or 0) and uses the affection status variable given in the pedigree file. It is recoded as 1 if the affection status from the pedigree file is 2 (affected), and zero otherwise (affection status is 1 - unaffected or 0 -unknown). The name of the default trait in

generated output is *affection*. With dichotomous traits, values can be changed using the *setafftrait* command.

7. FBAT-tools in Practice

The FBAT tools are illustrated at length in "A tour of FBAT. The provided documentation zooms in on particular genetic association problems and how they can be tackled with our FBAT-tools. Using real-life examples, guidance is provided on how to use the "FBAT-tools package" in different settings.

For a detailed description of the data sets used for illustration, we also refer to the FBAT web page. In particular, the data sets used for FBAT purposes are described in the appendix of "A tour of FBAT".

8. References

- Allison DB (1997). "Transmission-disequilibrium tests for quantitative traits." *Am J Hum Genet*, 60:676-690.
- Childhood Asthma Management Program Research Group (1999). The Childhood Asthma Management Program (CAMP): design, rationale, and methods. *Control Clin Trials*, 20:91-120.
- Blacker D, Haines JL, Rhodes L, Terwedow H, Go RCP, Harrell LE, Perry RT, Bassett SS, Chase G, Meyers D, Albert MS, Tanzi R (1997). "ApoE-4 and age at onset of Alzheimer's disease: the NIMH genetics initiative." *Neurology*, 48: 139-147.
- Xiao Ding, Christoph Lange, Xin Xu, Nan Laird. (2007) Family-based association tests with longitudinal measurements: a comparison of several approaches. Submitted.
- DeMeo DL, Lange C, Silverman EK, Senter JM, Drazen JM, Barth MJ, Laird NM, Weiss ST (2002). "Univariate and multivariate family-based association analysis of the IL-13 ARG130GLN polymorphism in the Childhood Asthma Management Program." *Genetic Epi*, 23: 335-348.
- Dempster AP, Laird NM and Rubin DB (1977). "Maximum likelihood from incomplete data via the EM algorithm." *Journal of the Royal Statistical Society*, Series B, 34:1-38.
- Fulker DW, Cherny SS, Sham PC, Hewit JK (1999). "Combined linkage and association sib-pair analysis for quantitative traits." *Am J Hum Genet*, 64:259-267.
- George V, Tiwari HK, Zhu X, Elston RC (1999). "A test of transmission/disequilibrium for quantitative traits in pedigree data, by multiple regression." *Am J Hum Genet*, 65:236-245.
- Heyde C (1997). "Quasi-likelihood and its application." Springer Series in Statistics.
- Horvath S. and Laird NM (1998) "A discordant-sibship test for disequilibrium linkage: No need for parental data." *Am J Hum Genet*, 63: 1886-1897.
- Horvath S, Xu X, Laird NM (2001). "The family based association test method: strategies for studying general genotype-trait associations*." Eu. J Hum Genet*, 9: 301-306.
- Horvath S, Xu X, Lake SL, Silverman EK, Weiss ST, Laird NM (2004) "Family based tests for association haplotypes with general trait data: application to asthma genetics." *Genet Epi*, 26(1): 61-69.
- Knapp M (1999). "A Note on Power Approximation for the Transmission/Disequilibrium Test." *Am J Hum Gen* 64: 1177-1185.
- Laird NM, Horvath S, Xu X (2000). "Implementing a unified approach to family-based tests of association*." Genet Epi*,19(suppl 1):S36-S42.
- Laird, N.M. and C. Lange, Family-based designs in the age of large-scale gene-association studies. Nat Rev Genet, 2006. 7(5): p. 385-94.
- Lake SL, Blacker D, Laird NM (2000). "Family-based tests of association in the presence of linkage*." Am J Hum Genet*, 67:1515-1525.

- Lange C, Laird NM (2002a). "Power calculations of a general class of family-based association tests: dichotomous traits." *Am J Hum Genet*, 71: 575-584.
- Lange C, DeMeo DL, Laird NM (2002b). "Power and design considerations for a general class of family-based association tests: quantitative traits." *Am J Hum Genet*, 71:1330-1341.
- Lange C, Laird NM (2002c). "On a general class of conditional tests for family-based association studies in genetics: the asymptotic distribution, the conditional power, and optimality considerations." *Genet Epi*, 23:165-180.
- Lange C, Silverman E, Weiss S, Xu X, Laird NM (2002d). "A Multivariate Family-Based Test using Generalized Estimating Equations: FBAT-GEE". *Biostatistics*, 1:1,1-15.
- Lange C, Blacker D, Laird NM (2004). "Family-based association tests for survival and times-to-onset analysis." *Statistics in Medicine*, 23: 179-189.
- Lange C, DeMeo D, Silverman EK, Weiss ST, Laird NM (2003a). "Using the noninformative families in family-based association tests: a powerful new testing strategy." *Am. J. Hum. Genet,* 73: 801-811.
- Lange C, Lyon H, DeMeo D, Raby B, Silverman EK, Weiss ST (2003b). "A new powerful non-parametric two-stage testing strategy for family-based association tests for testing multiple traits using all available data." *Am J Hum Genet,* 73: 801-811.
- Lange C, Silverman EK, Xu X, Weiss ST, Laird NM (2003c). "A multivariate transmission disequilibrium test." *Biostatistics*, 71: 195-206.
- Lazzeroni L, Lange K (2001). "A conditional inference framework for extending the transmission/disequilibrium test." *Hum Hered*, 48:67-81.
- Lewontin RC (1964). "The interaction of selection and linkage. I. General considerations; heterotic models." *Genetics* 49: 49-67.
- Liang K-Y, Zeger SL (1986). "Longitudinal data analysis using generalized linear models." *Biometrika*, 73:1,13-22.
- Lunetta KL, Farone SV, Biederman J and Laird NM (2000). "Family based tests of association and linkage using unaffected sibs, covariates and interactions." *Am J Hum Gen* 66: 605-614.
- Mokliatchouk O, Blacker D and Rabinowitz D (2001). "Association tests for traits with variable age at onset." Human Hederity, 51: 46-53.
- Press WM, Flannery BP, Teukolsky SA, Vetterling WT (1986). *Numerical recipes: The art of scientific computing*. New York, NY: Cambridge University
- Rabinowitz D (1997). "A transmission disequilibrium test for quantitative trait loci." *Hum Hered* 47: 342-350.
- Rabinowitz D, Laird NM (2000). "A unified approach to adjusting association tests for population admixture with arbitrary pedigree structure and arbitrary missing marker information." *Hum Hered*, 50:211-223.
- Rakovski C, Xu, X, Lazaras, R, and Laird N. (2007) "A new multimarker test for family-based association studies'. *Genet Epidemiol,* 2007. **31**(1): p. 9-17.
- Rakovski C, Xu, X and Laird N (2007) A New Minimum p-values Test for Multiple Markers. To appear in Human Heredity.
- Schneiter, K, Degnan, J, Cocoran, C and Nan Laird (2007) XFBATp: Exact Family-Based Association Tests. Submitted.

- Schaid, DJ (1996). “General score tests for associations of genetic markers with disease using cases and their parents.” *Genetic Epi,*13:423-49.
- Spielman RS, McGinnis RE, Ewens WJ (1993). “Transmission test for linkage disequilibrium: the insulin gene region and insulin-dependent diabetes mellitus (IDDM).” *Am J Hum Genet*, 65:578-580.
- Silverman EK, Kwiatkowski DJ, Sylvia JS, Lazarus R, Drazen JM, Lange C, Laird NM, Weiss S (2003). “Family-based association analysis of Beta-2 adrenergic receptor polymorphisms in the Childhood Asthma Management Program.” J Allergy Clin Immunol, Nov; 112(5):870-6.
- Tu IP, Balise RR and Whittemore AS (2000). “Detection of disease genes by use of family data II. Application to nuclear families.” *Am J Hum Gen,* 66:1341-1350.
- Whittaker J and Lewis C (1998). “The effect of family structure on linkage tests using allelic association.” *Am J Hum Gen,* 63:889-897.
- Xu, X., C. Rakovski, and NM Laird, An efficient family-based association test using multiple markers. *Genet Epidemiol,* 2006. **30**(7): p. 620-6
- Zhao H, Zhang S, Merikangas KR, Trixler M, Wildenauer DB, Sun F, Kidd KK (2000). “Transmission/Disequilibrium tests using multiple tightly linked markers.” *Am J Hum Genet*, 67:936-946.